Full Model Selection mit 15 unabhängigen Variablen: Ein Beispiel für SAS-Tuning bei komplexen rechenintensiven Aufgaben

Kaluscha, Rainer; Jankowiak, Silke; Krischak, Gert
Institut für Rehabilitationsmedizinische Forschung an der Universität Ulm
rainer.kaluscha@uni-ulm.de

Bei Rehabilitationsmaßnahmen der Rentenversicherung ist die anschließende berufliche Wiedereingliederung ein wichtiges Ziel. Dabei spielen aber nicht nur medizinische Parameter, sondern auch externe Einflussgrößen wie der Arbeitsmarkt eine Rolle. Zur Differenzierung dieser Effekte haben wir anonymisierte Routinedaten der Rentenversicherung zu 400.000 Rehabilitationsmaßnahmen [1] mit Arbeitsmarktdaten der Bundesagentur für Arbeit [2] zusammengeführt und versucht, mittels eines logistischen Regressionsmodells (SAS 9.3, Proc Logistic) die berufliche Wiedereingliederung vorherzusagen.

Da a priori nicht klar, welche Einflussgrößen tatsächlich relevant sind und wir auch an der Robustheit der Ergebnisse bei Variationen der Variablenselektion interessiert waren, haben wir uns entschlossen, eine full model selection durchzuführen. Bei 15 potientiell relevanten Confoundern waren also 2^15 = 32.767 Modelle zu prüfen und es stellte sich die Frage, wie Rechenzeit und Auswerteaufwand minimiert werden können; insbesondere, wenn z.B. aufgrund von Verfeinerungen der Stichprobe oder unterschiedlicher Operationalisierungen von Einfluss- oder Zielgrößen mehrere Durchläufe erforderlich sind.

Für die Auswertung stand uns ein leistungsfähiger Computeserver (16 CPUs, 256 GB RAM, SuSE Linux Enterprise 11, SAS 9.3) zur Verfügung. Der benötigte SAS-Code wurde mittels eigener Utilities aus einer Schablone generiert. Lässt man alle Modelle innerhalb eines Jobs berechnen, ergibt sich eine Gesamtlaufzeit von 39,5h. Durch Optimierungen (Parallelisierung auf 16 Jobs; Verlegung der Work-Library auf eine Ramdisk) ließ sich die Laufzeit um den Faktor 11,3 auf nur noch 3,5h verkürzen.

Anschließend wurden anhand der Modellgüte (c-Statistik, AIC) mit Hilfe von Unix-Utilities [3] die „besten“ Modelle ermittelt. Durch Vergleich dieser Modelle wurde geprüft, ob der Arbeitsmarkteinfluss generell relevant oder seine Bedeutung von der Variablen- oder Stichprobenselektion abhängig war. Diese Vorgehen vermeidet die Abhängigkeit der Ergebnisse von A-priori-Entscheidungen bei der Variablenselektion und kann so zur Gewinnung belastbarerer Aussagen beitragen.

Verweise:

[1] Forschungsdatenzentrum der Rentenversicherung: "Scientific Use File: Abgeschlossene Rehabilitation im Versicherungsverlauf 2002 - 2009 (SUFRSDLV09B)". Online: http://www.fdz-rv.de

[2] Bundesagentur für Arbeit: Aktuelle Daten - Arbeitslosigkeit und Grundsicherung für Arbeitsuchende nach Ländern. Online: http://statistik.arbeitsagentur.de

[3] Kaluscha R: Datenmanagement mit Oracle, SAS, Perl und Unix-Utilities: Werkzeuge für alle Fälle. KSFE 2007. Online: http://de.saswiki.org/images/7/77/11.KSFE-2007-Kaluscha-Datenmanagement-Werkzeuge.pdf